**RESUMOS ARTIGOS**

Amanda Ferreira de Souza

1. **Anunciando o WIT: um conjunto de dados de texto de imagem baseado na Wikipedia**

Este artigo se trata de um anúncio do conjunto de dados de texto de imagem baseado na Wikipédia (WIT), que é um grande conjunto de dados multimodal, criado pela extração de várias seleções de texto diferentes associadas a uma imagem de artigos da Wikipédia e links de imagem da Wikimedia. O que motivou a criação desse conjunto de dados foi a dificuldade encontrada originalmente em criá-los, pois eram feitos manualmente, ou extraídos de texto alternativo como legenda, porém isso limitava a quantidade de dados a ser criado, já que o processo manual é demorado e suscetível à falta de qualidade. Além disso, contava com outra dificuldade: a falta desse conjunto de dados em idiomas diferentes do inglês. A partir disso surgiu a pergunta “é possível superar essas limitações e criar um conjunto de dados multilíngue de alta qualidade e tamanho grande com uma variedade de conteúdo?” Portanto, foi criado um grande conjunto de dados multimodal, denominado WIT, que conta com curadoria de 37,6 milhões de exemplos de texto de imagem rica em entidades com 11,5 milhões de imagens exclusivas em 108 idiomas da Wikipédia. Dentre suas vantagens constam o tamanho do conjunto, que é o maior disponível publicamente; a multilinguagem, já que possui 10x ou mais idiomas do que qualquer outro conjunto de dados; a inclusão de muitas informações contextuais em nível de página e de seção, diferentemente dos outros conjuntos que oferecem apenas uma única legenda de texto (ou várias versões de uma legenda semelhante) para uma determinada imagem; e a riqueza em entidades do mundo real que são representadas no WIT.

Com essas vantagens, acredita-se que o conjunto de dados WIT ajudará os pesquisadores na construção de melhores modelos multimodais e na identificação de melhores técnicas de aprendizado e representação, levando a modelos aprimorados de aprendizado de máquina em tarefas do mundo real sobre dados visio-linguísticos.

1. **WIT: Conjunto de dados de texto de imagem baseado na Wikipedia para Aprendizado de máquina multilíngue multimodal**

Neste artigo, foram apresentados Conjunto de dados de texto de imagem baseado na Wikipédia (WIT), para facilitar a aprendizagem multilíngue multimodal, sendo que as técnicas de modelagem multimodal visam alavancar grandes conjuntos de dados visio-linguísticos de alta qualidade para aprender informações complementares (entre imagens e modalidades de texto).

O WIT é composto por um conjunto com curadoria de 37,6 milhões de exemplos de texto de imagem rico em entidade com 11,5 milhões de imagens em 108 idiomas da Wikipédia. Seu tamanho permite que o WIT seja usado como um conjunto de dados de pré-treinamento para modelos multimodais.

Vantagens do WIT: Primeiro, o WIT é o maior conjunto de dados multimodal pelo número de exemplos de texto de imagem por 3x (em momento da escrita). Em segundo lugar, o WIT é massivamente multilíngue (primeiro de seu tipo) com cobertura em mais de 100 idiomas (cada um com pelo menos 12 mil exemplos) e fornece textos multilíngues para muitas imagens. Terceiro, WIT representa um conjunto mais diversificado de conceitos e entidades mundiais em relação ao que os conjuntos de dados anteriores cobrem. Por último, WIT fornece um conjunto de testes do mundo real muito desafiador, como ilustrado empiricamente usando uma tarefa de recuperação de texto de imagem como exemplo.

O WIT é criado pela extração de vários textos associados a uma imagem (por exemplo, a descrição de referência) de artigos da Wikipédia e links de imagens da Wikimedia. Isso foi acompanhado por uma filtragem rigorosa para reter apenas altas associações de qualidade de imagem-texto

Começou-se com todas as páginas de conteúdo da Wikipédia (ou seja, ignorando outras páginas que tenham discussões, comentários e afins). Esses números têm cerca de 124 milhões de páginas em 279 idiomas. Foi usado um pipeline Flume para processar programaticamente, filtrar, limpar e armazenar os dados da Wikipedia. Em seguida, imagens e diferentes textos relacionados à imagem junto com alguns metadados contextuais (como o URL da página, título da página, descrição…) foram extraídas. Isso rendeu cerca de 150 milhões de tuplas de dados de imagem, dados de texto, dados contextuais, que foram a entrada para os diferentes filtros descritos nas seções subsequentes.

Sobre os textos que descrevem as imagens: Eles vêm de várias fontes diferentes. Os três diretamente associados à imagem são:

(1) Descrição da referência (abreviada como ref ):

(2) Descrição da atribuição (abreviada como attr):

(3) Descrição do texto alternativo (abreviado como alt):

Ao longo do estudo foram feitas diversas mudanças como limpeza dos textos com pouca informação, foram feitos filtros nas imagens das tuplas assim como uma filtragem adicional para garantir um conjunto de dados de alta qualidade livre de conteúdo impróprio.

Como dito anteriormente, o conjunto de dados resultante é significativamente maior que os anteriores com mais de 37 milhões de tuplas (imagem, texto(s), contexto), abrangendo 108 idiomas e cobrindo 11,5 milhões de imagens únicas. Entre seus muitos aspectos únicos e inéditos:

• Vários textos por imagem: Mais da metade das tuplas (19,4M) têm dois ou mais textos por imagem de referência, atribuição e textos alternativos.

• Altamente multilíngue: Quase metade dos mais de 100 idiomas, contêm mais de 100 mil tuplas de texto de imagem exclusivas e mais de 100 mil tuplas exclusivas imagens.

• Grande cobertura multilíngue: WIT pode ser usado para gerar mais de 50 milhões de pares de línguas cruzadas (ou seja, texto descrições em diferentes idiomas para a mesma imagem) de 3,1 milhões de imagens diferentes usando apenas a referência e os textos alternativos.

• Compreensão contextual: WIT também é o primeiro conjunto de dados, fornecendo para a compreensão de legendas de imagem no contexto da página e texto ao redor.

Como resultado, foi provado que é adequado para uso de diversas maneiras, incluindo pré-treinamento de modelos multimodais, ajuste fino nos modelos de recuperação de imagem-texto ou construção de representações multilíngues para nomear alguns. Com a análise e avaliação feita, foi validado que o WIT é um conjunto de dados de alta qualidade com texto de imagem de forte alinhamento.Também foi demonstrado empiricamente o uso deste conjunto de dados como um conjunto de pré-treinamento e ajuste fino, e no processo foi descoberto algumas deficiências dos conjuntos de dados existentes.

1. **Desenvolvimento de banco de dados de imagens de treinamento usando Web Crawling para monitoramento de canteiro de obras baseado em modelo visão.**

Esse artigo começa nos informando que propõe uma estrutura automatizada que constrói um banco de dados de treinamento grande e de alta qualidade para construção de monitoramento de canteiro de obras. O monitoramento baseado em visão em canteiros de obras é uma das tarefas mais críticas para o gerenciamento de projetos bem-sucedidos. Com ele é possível reconhecer, por exemplo, alguns trabalhadores que não usam roupas, equipamentos de proteção (por exemplo, capacetes) ou que estão mostrando comportamentos perigosos (por exemplo, acessar zonas onde o equipamento está sendo usado). É também possível auxiliar os gerentes de projeto a realizarem treinamento de segurança para eliminar esse potencial risco e prevenir acidentes. Além disso, o monitoramento de canteiros de obras permite identificar os tipos de atividade de recursos de construção (por exemplo, trabalhando ou ocioso) e medir a quantidade de trabalho concluído em um determinado tempo (por exemplo, volume de solo escavado por hora). Para isso ser possível, a estrutura automatizada consiste em três processos principais: (1) coleta automatizada de imagens de construção usando Web Crawling (ferramenta que pesquisa imagens em sites de compartilhamento de dados online e coleta as imagens com base no HTML dos sites.) (2) rotulagem de imagem automatizada usando um modelo de segmentação de imagem e (3) sobreamostragem cruzada foreground-background totalmente aleatória.

(1) Coleta automatizada de imagens de construção usando rastreamento na web

Objetivo: coletar automaticamente as numerosas imagens de construção que estão espalhadas pelos websites. Essas imagens representam diversas características visuais dos canteiros de obras, como tipos de recursos de construção (por exemplo, caminhões basculantes, escavadeiras, carregadeiras, trabalhadores) e operações (por exemplo, carga, despejo, inatividade). Para isso foram realizados 3 passos.

Primeiro, os autores desenvolveram um dicionário de palavras-chave que pode ser usado para efetivamente pesquisar imagens de sites de compartilhamento de dados online. Um dicionário de palavras-chave pode abranger imagens a serem investigadas por motores de busca e desempenham um papel na filtragem de imagens que não se encaixam no objetivo do estudo.

Por exemplo, alguém pode pesquisar a palavra-chave “escavadeira” para coletar imagens de escavadeiras, porém imagens desnecessárias que não são necessárias para o treinamento podem aparecer, como as imagens de peças de escavadeiras ou manuais de escavadeira. Assim, os autores desenvolveram um dicionário de palavras-chave específico de construção adicionando os nomes de empresas de manufatura para a palavra-chave “escavadeira”, como escavadeira Caterpillar, escavadeira Komatsu, escavadeira Volvo e escavadeira Kubota. Além disso, algumas palavras foram adicionadas e usadas na pesquisa para derivar resultados de pesquisa adequados para a finalidade do estudo. Por exemplo, um dos excelentes sites de compartilhamento online do Google tem a função de usar o sinal “-” para remover buscas desnecessárias dos resultados. Além disso, tem função adicional para colocar uma palavra entre as aspas duplas (por exemplo, "caminhão basculante") para obter os resultados que incluem a palavra exata. Para busca por palavra-chave, “um” e “único” foram adicionadas na frente das palavras citadas no dicionário de palavras-chave, e “-brinquedo”,“-adesivo”, “-desenho técnico" e “-ícone” foram adicionados atrás das palavras no dicionário de palavras-chave. Assim, as imagens nas quais existe apenas um objeto poderiam ser coletadas e imagens de brinquedos ou adesivos podem ser excluídas.

Em segundo lugar, as palavras-chave incluídas no dicionário de palavras-chave são alimentadas em um Web Crawling, que pode pesquisar e baixar automaticamente imagens relacionadas a palavras-chave na Internet, e somente as URLs das imagens são extraídas. Por último, as imagens são coletadas usando os URLs que foram coletados anteriormente.

(2) Rotulagem de imagem automatizada usando um modelo de segmentação de imagem pré-treinado

Este processo anota automaticamente as informações do rótulo dos recursos da construção (ou seja, tipos e locais) nas imagens coletadas no processo anterior. Para fazer isso, os autores exploraram um modelo de segmentação semântica que foi pré-treinado por um

conjunto de dados de referência disponível (ou seja, conjunto de dados MS COCO). Sabe-se que este modelo de segmentação pode distinguir o primeiro plano e o plano de fundo das imagens de construção coletadas, mesmo que não haja nenhum domínio específico nos dados de treinamento fornecidos. Assim, é possível posicionar recursos de construção nas imagens usando as informações do primeiro plano e dos pixels restantes,

e rotular seus tipos pelas palavras-chave usadas para a pesquisa na Internet. Em alguns casos, imagens obtidas pelo Web Crawling que foram tiradas muito perto de um objeto, pode dificultar a determinação do tipo do objeto capturado. Essas imagens não são necessárias para o desenvolvimento do modelo porque não representam as características visuais gerais do objeto. Observando que a proporção do fundo de tais imagens seria relativamente baixa (por exemplo, 10%) em comparação com outras imagens (por exemplo, 50%), eles foram excluídos do banco de dados de treinamento do estudo. Além disso, poucas imagens com mais de um objeto também podem ser coletadas e devem ser excluídas. Para isso, os autores experimentaram uma série de diferentes escalas de objetos para encontrar o limite da proporção de fundo apropriada. O limite mínimo da razão de fundo foi definido para ser 40%, enquanto o limite máximo foi fixado em 70%. Com base nos resultados da segmentação semântica das imagens, as localizações em nível de pixel de recursos de construção nas imagens, ou seja, as coordenadas *x* e *y*, podem ser identificadas. Os dados de anotação [object\_type, xmin, ymin, xmax, ymax] são eventualmente produzidos automaticamente para todas as imagens coletadas.

(3) Amostragem cruzada de foreground-background totalmente aleatória

Nesse processo, as imagens que foram coletadas e rotuladas nas etapas anteriores são combinadas com as imagens de fundo dos canteiros de construção, a fim de desenvolver um banco de dados com imagens de treinamento mais foto realista. As imagens coletadas e rotuladas têm aparências visuais e formas diversas, e os objetos se localizam no meio das imagens na maioria dos casos. No entanto, os recursos de construção reais, como equipamentos pesados, são todos diferentes em tamanho e distância da câmera, então os tamanhos dos recursos de construção projetados na imagem também são diferentes. Assim, se as imagens coletadas e rotuladas forem usadas diretamente para treinar um modelo para detectar objetos de construção, as várias informações sobre seus tamanhos e localizações não podem ser treinadas, e o desempenho pode ser reduzido.

Os autores, assim, desenvolveram um processo de sobreamostragem cruzada de três etapas para minimizar as discrepâncias visuais entre as imagens de construção rastreada da web e as imagens de construção do mundo real. Para isso, objetos são extraídos das imagens coletadas com base nas informações de rotulagem, e então eles são sintetizados nas imagens reais do site. Para ser específico, os recursos de construção são primeiro cortados das imagens coletados, usando informações de rotulagem (ou seja, locais de recursos). Subsequentemente, os tamanhos dos recursos de construção cultivados são

redimensionado, mantendo as proporções dos recursos para composição em imagens reais do site. Por último, os recursos redimensionados são sobrepostas em imagens do mundo real do canteiro de obras. Nesse passo, os recursos no local são normalmente ligados ao solo, então os recursos redimensionados são compostos aleatoriamente no restante da imagem, exceto sua parte superior.

Por meio desse processo, é possível criar rapidamente imagens que tenham o visual de primeiro e segundo plano com semelhantes combinações como as capturadas nos próprios canteiros de obras. Isso também aumenta o número de imagens de treinamento porque os recursos extraídos (primeiro plano) e as imagens do canteiro de obras (fundo) são combinadas. Usando o banco de dados de imagem de treinamento desenvolvido, seria

possível sintetizar dados mais fotorrealistas e ensinar tanto positivas (por exemplo, escavadeira, caminhão basculante) e negativas (por exemplo, ambientes circundantes)

instâncias ao modelo.

Assim, os autores referiram-se a este processo como sobreamostragem cruzada de primeiro e segundo plano, porque cria inúmeras imagens de treinamento combinando recursos (foreground) coletados da Internet e do site real (background).

Para validar o framework proposto, três experimentos diferentes foram realizados: (1) comparação das imagens coletadas com e sem o uso do dicionário de palavras-chave, (2) comparação de rotulagem por um modelo de segmentação semântica pré-treinado e rotulagem por uma pessoa, e (3) comparação dos desempenhos de dois monitoramentos baseados em visão modelos desenvolvidos pelo treinamento de banco de dados de imagem com e sem uso de sobreamostragem cruzada foreground-background.

Com isso, foi possível construir um banco de dados de treinamento, composto por 5864 imagens, para detecção de objetos de construção em 53,5 min. O modelo de aprendizado profundo treinado pelo banco de dados detectou com sucesso recursos de construção com uma precisão média de 92,71% e uma taxa de reconvocação de 88,14%. Os resultados deste estudo podem reduzir o tempo e o esforço necessários para desenvolver tecnologias de monitoramento de site baseadas em visão.